OptCaching: A Stackelberg Game and Belief Propagation Based Caching Scheme for Joint Utility Optimization in Fog Computing

Kai Lei*, Yingying Xie*, Jian Shi*,Haijun Zhang†,Gong Zhang‡,Bo Bai‡
*Shenzhen Key Lab for Information Centric Networking & BlockChain Technology (ICNLAB),
School of Electronics and Computer Engineering (SECE), Peking University, Shenzhen 518055, P.R. China
Email: leik@pkusz.edu.cn, 1701213646@sz.pku.edu.cn, 1501213960@sz.pku.edu.cn
†University of Science and Technology Beijing, Beijing 10083, P.R. China
Email: haijunzhang@ieee.org
‡Huawei Future Network Lab, Hongkong, China
Email: nicholas.zhang@huawei.com, Corresponding Author: baibo8@huawei.com

***Abstract*—Fog Computing which extends the cloud computing paradigm to the edge of the network provides great opportunities for applications with stringent latency requirement. How to allocate the limited caching resources of Fog Nodes (FNs) influences the performance of the fog computing system. In contrast to previous works on caching resource allocation with users' utility as the only consideration, we propose OptCaching which jointly optimize the utility of all network participants including Content Provider (CP), Internet Service Provider (ISP) and users. With caching incentive introduced, utility functions of these three roles are defined. Our joint utility optimization caching scheme is conducted in two stages combining global and local decision making. Firstly, interaction between CP and ISP is modeled as a non-cooperative hierarchy Stackelberg game to make decision on incentive caching prices and global caching amount aiming at optimizing the utility of all network participants. Secondly, for the purpose of further optimizing the utility of users, a belief propagation based cache placement algorithm which utilizes global caching amount constraint and local information is conducted by FNs to reduce users' average download delay. Mathematical analysis and simulation results show that the utility of CP, ISP and users are jointly optimized at Stackelberg equilibrium. The utility of users is further optimized by belief propagation based cache placement algorithm with users' average download delay reduced by 33.7% compared with global popularity based caching strategy.**

***Index Terms*—cache, utility optimization, fog computing, Stackelberg game, belief propagation**

I. Introduction

Due to the remote deployment of large scale data centers, existing cloud-based application frameworks face issues such as high service latency, network overhead, I/O bottleneck, etc [1]. Fog computing [2] has been proposed as a novel computing architecture which allows applications to fully utilize free computing, storage and network resources of edge and end devices. Compared with cloud computing, the advantages of fog computing consists in relieving the transmission burden of backbone network, reducing the service delay of users, potential in providing location-aware services, etc. Many promising applications, e.g., virtual reality (VR) and augmented reality (AR) with low latency and large scale content distribution demands, can profit from fog computing. For example, leveraging fog computing's advances in real-time computing response and proximal storage, VR games like the one in Steven Allan Spielberg's film "Ready Player One" may become reality soon.

Considering a typical fog computing scenario with 3 types of network participants including CP, ISP and user as shown in Fig.1. We refer FNs to network infrastructures such as base station, wireless access point, etc, which are under the management of ISP and have a certain amount of available caching resources. The introduction of cache in fog computing is benefitial to all network participants. CP can lease the caching resources of FNs from ISP to reduce its service delay to users and retransmission cost of duplicate requests. ISP profits from caching incentive provided by CP and alleviated backbone traffic pressure. Users enjoy both reduced transmission delay and transmission fee. Due to the scarcity of caching resources, the diversity and dynamicity of resource requirements and the influence of caching resource allocation on the benefits of all network participants, it is worth to study how to allocate the limited caching resources rationally in fog computing scenario. Specifically, how to determine the caching incentive, how much content should be cached by ISP and what is the reasonable FN-content match.

We consider a continuous caching decision making process which accurately models the real interaction in fog computing [3]. For a more intuitive understanding, take Youku as CP and China Mobile as ISP for example. After Youku offers incentive caching prices, China Mobile makes decision on global caching amount according to the provided incentive. Then the cache placement problem is solved by FNs for its convenient and fully awareness of the requests distribution of

This work has been financially supported by National Major Science and Technology Infrastruture: CENI (National Development and Reform Commission (NDRC) [2016] 2533) and Shenzhen Key Fundamental Research Project (No: JCYJ20170412151008290 and JCYJ20170412150946024).

local users. This continuous caching decision making process is ought to be conducted periodically, for example off-peak time at midnight everyday, to refresh the caching content of FNs according to the altered requests distribution.

Utility function encodes the benefit derived by a network participant in this considered content caching [4]. With caching incentive introduced, the utilities of CP, ISP and users are defined separately. Previous works on caching resource allocation in fog computing focused on reducing the utility of one certain network participant while leaving the utilities of others undiscussed [5] [6] [7]. Since CP, ISP and users all can benefit from the caching of fog computing, they may be selfish and tend to maximize their own utility. In order to encourage cooperation and enhance the overall performance of the fog computing system, we claim that the utilities of all network participants are supposed to be considered in the design of caching scheme. Therefore, we investigate a joint utility optimization caching scheme.

Capturing the nature of continuous decision making process of our considered scenario, our joint utility optimization approach is conducted in two stages. In the first stage, the utilities of CP, ISP and users are jointly optimized. Specifically, CP decides the incentive caching price and then jointly considering the incentive price provided by CP and the requests distribution of users, ISP makes decision on global caching amount. In the second stage, each FN decides which content to be cached to further optimize the utility of users in a distributed way. The main contributions of this paper lies in:

- Combining incentive scheme, we formulate the interaction between CP and ISP as a Stackelberg game model to optimize the utilities of all network participants. The existence of Stackelberg equilibrium is proved by mathematical analysis.
- A belief propagation based distributed cache placement algorithm is proposed to further optimize the utility of users. Simulation results show its rapid convergence speed and effectiveness in reducing average download delay of users.
- Simulation results show that the utility of CP, ISP and users are jointly optimized at Stackelberg equilibrium. The pricing behavior of CP and caching behavior of ISP are effectively shaped by our incentive caching scheme.

The remainder of this paper is structured as follows. Section II briefly summarizes the related works with respect to resource pricing and allocation. Section III presents the 3-layer fog computing network model. In Section IV, we formulate the caching resource allocation problem. In Section V, a joint utility optimization caching resource allocation strategy separated in two stages is proposed. Then, Section VI demonstrates the simulation results and analysis. We conclude this paper in Section VII.

II. RELATED WORK

Many research effort has been dedicated to the resource pricing and allocation problem. Facing the information asymmetry problem of users' true valuations of content and requirements of delivery quality, [7] proposed an auction mechanism to derive the optimal caching scheme from the perspective of the service providers (SPs). With a hierarchical mobile edge computing architecture introduced, [8] separated the computing and communication resources allocation in two time scale and formulated a binary linear programming (BLP) aiming at maximizing the profit of the service provider. The average profit of the network service provider (NSP) and video retailers (VRs) in a small-cell video caching system are jointly optimized via a Stackelberg game approach in [9], where uniform and non-uniform pricing schemes are compared in terms of reducing backhaul costs and sum profit of NSP and VRs. [10] investigated the optimal strategy of assigning files to edge caches in coded case and uncoded case separately for the purpose of minimizing the average download delay. Our work distinguishes from those previous works because of the joint optimization of the utilities of CP, ISP and users.

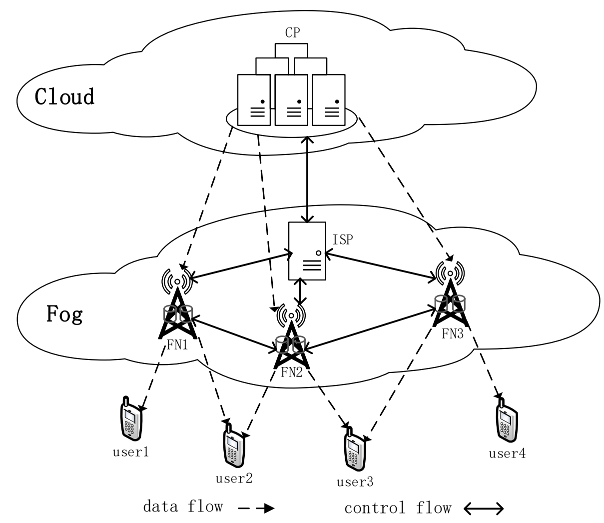


Fig. 1. 3-layer fog computing architecture

III. NETWORK MODEL

Considering a fog computing scenario, as shown in Fig.1. Suppose there are one CP, one ISP, L FNs, denoted as $\Omega = \{n_1, ..., n_L\}$ and G users, denotend as $U = \{u_1, ..., u_G\}$ respectively. Generally, FN refers to network infrasturcture such as base station and access point which is equipped with a limited amount of cache and has a certain coverage range. User u_g can be served by multiple FNs if he locates in the overlapping coverage range. Otherwise, he is only served by one FN. As for user u_g, its potential serving FN is denoted as $\Omega_g = \{n_l \in \Omega\}$. Let $U_l = \{u_g \in U\}$ denotes the users served by FN n_l. The topology of pysical transmission network between CP and FNs, FNs and users follows Second Topology [11], thus the impact of network congestion is not considered in this paper.

Data is segmented into equal-sized chunks and users send chunk-level requests. In this proposed incentive caching scheme, ISP receives reward from CP for caching data. Due

TABLE I
LIST OF NOTATIONS

Symbol	Meaning
$\Omega = \{n_1, ..., n_L\}$	The set of FNs
$U = \{u_1, ..., u_G\}$	The set of users
$F = \{f_1, f_2, ..., f_K\}$	The data class set
Ω_g	Potential serving FNs for user u_g
U_l	Users can be served by FN n_l
S_F	Total number of requests
λ_k	The number of requests for data f_k
ω_k	The percentage of requests for data f_k from all users
p_{gk}	The percentage of requests for data f_k from user u_g
$V = \{v_1, v_2, ..., v_K\}$	The incentive caching price vector
$X = \{x_1, x_2, ..., x_K\}$	The global caching amount vector
$H = \{h_{lk}\}$	The cache placement binary variables
A_l	The cache capacity of FN
v_{con}	Unit content price of CP
v_{tra}	Unit transmission price of ISP
g	The weight of users' dissatisfaction in the utility of CP
C	Unit transmission cost of ISP
C_0	Unit caching cost of ISP
θ	The increasing marginal cost parameter of ISP
ζ_{gk}	The income of user g_k by consuming data f_k
$\overline{D}_{gk}(H)$	The average download delay of user g_k requesting f_k
s_i	the ith variable node
F_j	the jth function node
Γ_i^s	Neighboring function nodes of variable node s_i
Γ_j^F	Neighboring variable nodes of function node F_j

to the numerous amount of data, it is impossible for CP to determine incentive prices in chunk-level. Therefore, We divide data into K classes according to their popularity and determine incentive price for each class of data. The data class set is denoted by $F = \{f_1, ..., f_K\}$. Let λ_k denotes the amount of data f_k that users requests and S_F denotes the total amount of users requests for all classes of data. The propotion of λ_k in S_F is denoted as w_k, thus $w_k = \frac{\lambda_k}{S_F}$. p_{gk} represents the percentage of requests for data f_k of user u_g in the total requests of user u_g for all classes of data. And we have $\sum_{k=1}^{K} p_{gk} = 1$. Let $V = \{v_1, ..., v_K\}$ denotes the caching price offered by CP. The caching amount of each class of content in ISP is denoted as $X = \{x_1, ..., x_K\}$. We also suppose that each FN has a finite storage capacity of A_l which means each FN can cache A_l chunks of data at most. The binary variable h_{lk} indicates whether a chunk of data f_k is cached at FN n_l. That is, $h_{lk} = 1$ if a chunk of data f_k is stored in the buffer of FN n_l, otherwise $h_{lk} = 0$. Therefore, the global caching amount vector $X = \{x_k\}(k \in [1, K])$ and cache placement decision matrix $H = \{h_{lk}\}(l \in [1, L], k \in [1, K])$ represents the caching allocation strategy. A list of notation is summarized in Table I.

A. Utility of CP

$$W_s(V)=\sum_{k=1}^{K}\Big\{v_{con}\lambda_k-v_kx_k-v_{tra}(\lambda_k-x_k)-g\big[1-S_k\big]\Big\} \quad (1)$$

$$S_k = \begin{cases} \frac{-\frac{1}{S_F}x_k^2+2w_kx_k}{w_k\lambda_k}, & x_k < \lambda_k \\ 1, \quad x_k \geqslant \lambda_k \end{cases} \quad (2)$$

CP gets profit by charging users for consuming its content. The unit price is set as v_{con}, thus the income of CP equals to $\sum_{k=1}^{K} v_{con}\lambda_k$. CP pays ISP for the caching service and transmission service. With ISP's unit transmission fee denoted as v_{tra}, the transmission fee paid to ISP is proportional to the amount of cache miss that is $\sum_{k=1}^{K} v_{tra}(\lambda_k - x_k)$. CP can set a lower unit price v_k in order to decrease its cost, under which circumstance, ISP tends to cache fewer content and therefore users experience longer download delay. From a long-term perspective, users may not consume this CP's content any more. Therefore, it is reasonable to consider user's dissatisfaction level as part of CP's cost. Let g denotes the weight of users' dissatisfaction level in the utility of CP. As shown in (2), users satisfaction level is modeled as a nondecreasing function S_k whose first derivative is nonincreasing. The design philosophy of S_k lies in that users are more satisfied with larger amount of cache while get less sensitive as cache amount increases [12]. S_k is normalized into the interval $[0, 1]$. In the case of $x_k \geqslant \lambda_k$, namely, ISP caches adequate amount of data f_k, user's satisfaction get saturated and equals to one. Noticed that there is an inherent conflict between incentive caching price offered by CP and user's dissatisfaction level. The utility of CP is influenced by the incentive caching prices V and global caching amount X.

B. Utility of ISP

$$W_d(X)=\sum_{k=1}^{K}\Big\{v_{tra}(2\lambda_k-x_k)+v_kx_k-C(2\lambda_k-x_k)-\delta_k\Big\} \quad (3)$$

$$\delta_k = C_0x_k(\frac{x_k}{\lambda_k})^{\theta} \quad (4)$$

ISP's income consists of transmission fees from users and CP and caching reward from CP. The transmission payment from users and CP equals to $\sum_{k=1}^{K} v_{tra}\lambda_k$ and $\sum_{k=1}^{K} v_{tra}(\lambda_k - x_k)$ respectively. ISP's caching reward equals to CP's caching cost as described in the Utility of CP. As for ISP, suppose the unit caching cost is C_0 and unit transmission cost is C. In the case of cache miss, ISP has to forward the user requests to CP, resulting in transmission cost to ISP which is proportional to the amount of cache miss content $(\lambda_k - x_k)$. With addition to transmitting λ_k chunks of data f_k to users, the total transmission cost of ISP sums up to $\sum_{k=1}^{K} C(2\lambda_k - x_k)$. Let δ_k denotes the caching cost of ISP for caching x_k chunks of data f_k. δ_k increases as x_k increases. Especially, δ_k equals to zero when x_k equals to zero. Therefore, the caching cost is defined as (4), where the form of power law cost function reflects the general rule of increasing marginal cost. θ denotes the increasing marginal cost parameter of ISP. The utility of ISP is also influenced by the incentive caching prices V and global caching amount X like the utility of CP does.

C. Utility of users

$$W_U(H)=\sum_{g=1}^{G}\sum_{k=1}^{K}\Big\{\zeta_{gk}-p_{gk}\overline{D}_{gk}(H)\Big\}-\sum_{k=1}^{K}\lambda_k(v_{con}+v_{tra}) \quad (5)$$

Variable ζ_{gk} denotes the mental satisfaction of user u_g gained by consuming the requested data f_k, which is out of the discussion of this paper. Users pay content fee to CP and transmission fee to ISP which are fixed since CP and ISP' pricing and users' content demand are determined. Moreover, users' download delay is considered as part of users' cost with $\overline{D}_{gk}(H)$ denoting the average download delay for user u_g to download data f_k. Then the normalized average download delay equals to $p_{gk}\overline{D}_{gk}(H)$. $\overline{D}_{gk}(H)$ is influenced by the global caching amount and the specific cache placement strategy. Hence, a rational caching scheme is required as the only way to maximize the utility of users.

IV. Problem Formulation

In the fog computing scenario, cache pricing decisions of CP and global caching amount decisions of ISP exert enormous influence on the utilities of all network participants, which is exemplified as follows. As for CP or ISP, they desire to maximize their own utility. To achieve this purpose, CP may set a lower caching incentive price to cut down costs. In the perspective of ISP, it tends to cache less content for CP since the provided caching incentive is not profitable. As a result, users' requests are mainly forwarded to CP and longer download delay is incurred, in which case, ISP bear higher transmission cost, users suffer from longer download delay and CP may face lose of users. None of ISP, CP or users obtain optimal utility. Therefore, we can conclude that CP's incentive caching price and ISP's cache allocation have cross impact on the utility of each other. Besides, the decision on global caching amount of ISP influences the utility of users as we demonstrated above. Hence, rational decisions on cache pricing and global caching amount are of great importance to optimizing the utilities of all network participants.

The optimization problem of CP is formulated as (6) and (7). In the perspective of CP, it attempts to maximize its utility by offering a reasonable incentive caching price under the constraints presented in (7). The first constraint restricts the incentive caching price in the range of $[0, v_{high}]$. When offered the caching incentive price as v_{high}, any ISP is willing to provide caching service. Theoretically, CP can set v_k as any non-negative number, but set v_k higher than v_{high} is not more profitable for CP. The weight of user's dissatisfaction level in the utility of CP is set as equal or greater than zero. The third constraint means the utility of CP should be greater than zero.

$$\max_{V} W_s(V) \tag{6}$$

$$s.t. \begin{cases} 0 \leqslant v_k \leqslant v_{high} \\ g \geqslant 0 \\ v_{con}\lambda_k \geqslant v_k x_k + v_{tra}(\lambda_k - x_k) + g[1 - S_k] \end{cases} \tag{7}$$

As for ISP, its optimization problem is presented as formula (8) and (9). Under several constraints, the objection of ISP is to maximize its utility by making a decision on the global caching amount for various classes of data given the incentive caching prices offered by CP. The first constraint means ISP can not provide caching that exceeds its cache capacity. Since caching more content than users' expectation is a waste of the scarce caching resources and results in no utility to any network participants, we constraint x_k in the interval of $[0, \lambda_k]$. Similarly, the last constraint means the utility of ISP is a non-negative number.

$$\max_{X} W_d(X|V) \tag{8}$$

$$s.t. \begin{cases} \sum_{k=1}^{K} x_k \leqslant \sum_{l=1}^{L} A_l \\ 0 \leqslant x_k \leqslant \lambda_k \\ v_{tra}(2\lambda_k - x_k) + v_k x_k \geqslant C(\lambda_k - x_k) + \delta_k \end{cases} \tag{9}$$

As for users, in order to maximize its utility, one approach is to increase income, and the other is to reduce costs. As described above, the discussion of users' income by consuming the content is out of this paper's scope. The payment to ISP and CP is also unchangeable as users' requests are fixed. The download delay is related to global caching amount and cache placement strategy. Global caching amount is solved by ISP while FNs settle the cache placement problem. Different cache placement strategy results in different average download delay to users. Therefore, the only way for FNs to optimize utility of users is to decrease average download delay. We define the objective function of cache placement strategy as formula (10) and (11), i.e., finding the optimal H under certain constraints. The first constraint declares FN's caching capacity. The second one means the amount of content replicates of f_k cached by FNs should not exceed the amount decided by ISP in the incentive caching model. The last constraint guarantees h_{lk} to be a binary variable.

$$\min_{H} \frac{1}{G} \sum_{g=1}^{G} \sum_{k=1}^{K} p_{gk} \overline{D}_{gk}(H) \tag{10}$$

$$s.t. \begin{cases} \sum_{k=1}^{K} h_{lk} \leqslant A_l, \quad \forall n_l \in \Omega \\ \sum_{l=1}^{L} h_{lk} \leqslant x_k, \quad \forall f_k \in F \\ h_{lk} \in \{0, 1\}, \quad \forall f_k \in F, n_l \in \Omega \end{cases} \tag{11}$$

V. System Analysis

After formulating the utility maximization problem of CP, ISP and users, we present our joint utility optimization caching scheme in this section. Following the continuous decision making nature of the interactions among CP, ISP and users, the scheme is separated into two stages. Interaction between ISP and CP makes global caching decisions while interaction between users and FNs operates in a distributed manner with local information utilized only to make cache placement decisions.

A. Interaction between ISP and CP

In this section, we consider a continuous decision making process between CP and ISP: firstly, jointly considering the content popularity distribution and ISP's possible reaction, CP decides the caching incentive price $V = \{v_1, ..., v_K\}$ to optimize its utility. Given the information of user's requests distribution and caching incentive price offered by CP, ISP decides the amount of each class of content to be cached $X=\{x_1,...,x_K\}$ in order to maximize its utility. The decision of ISP depends on the decision of CP while the decision of CP is made by estimating the possible reaction of ISP, which is perfectly in accordance with Stackelberg game model. Based on the above consideration, our incentive caching model is formulated as a non-cooperative hierarchy Stackelberg game where CP acts as leader and ISP acts as follower. The solution of our incentive caching model is inspired by [13].

Second order derivatives the utility of ISP with respect to x_k:

$$\frac{\partial^2 W_d}{\partial x_k \partial x_h} = \begin{cases} -\frac{C_0\theta(\theta+1)}{\lambda_k^{\theta}}(\frac{x_k}{\lambda_k})^{\theta-2}, & k=h \\ 0, & k\neq h \end{cases} \tag{12}$$

If $\theta>0$, the Hessian matrix is negative definite matrix [14]. Then, the max utility of ISP W_d is obtained at setting the first order derivative of $W_d(X)$ with respect to x_k as 0 which is denoted as x_k^*. We can conclude that given the caching incentive price $V=\{v_1,...,v_K\}$ and having $\theta>0$, there exists a solution of $X^*=\{x_1^*,...,x_k^*\}$ that maximizes the utility of ISP. And we have

$$x_k^* = \lambda_k\left[\frac{v_k+C-v_{tra}}{C_0(\theta+1)}\right]^{\frac{1}{\theta}} \tag{13}$$

Replacing x_k in $W_s(V)$ with x_k^*, we obtain $W_s(V)$ as a function of v_k and the second partial derivative of $W_s(V)$ respect to v_k is shown as follows:

$$\frac{\partial^2 W_s}{\partial v_k \partial v_h} = \begin{cases} -2\frac{\partial x_k^*}{\partial v_k} - \frac{2}{\lambda_k}(\frac{\partial x_k^*}{\partial v_k})^2 + U_{mk}\frac{\partial^2 x_k^*}{\partial v_k^2}, & k=h \\ 0, & k\neq h \end{cases} \tag{14}$$

$$U_{mk} = v_{tra}+g-v_k+2(1-\frac{x_k}{\lambda_k}) \tag{15}$$

As shown in equation (14), the second partial derivative is related to the first and second derivative of x_k^* as shown below:

$$\frac{\partial x_k^*}{\partial v_k} = \frac{\lambda_k}{\theta[C_0(\theta+1)]^{\frac{1}{\theta}}}\left(v_k+C-v_{tra}\right)^{\frac{1}{\theta}-1} \tag{16}$$

$$\frac{\partial^2 x_k^*}{\partial v_k \partial v_h} = \begin{cases} \frac{\lambda_k(\frac{1}{\theta}-1)}{\theta[C_0(\theta+1)]^{\frac{1}{\theta}}}\left(v_k+C-v_{tra}\right)^{\frac{1}{\theta}-2}, & k=h \\ 0, & k\neq h \end{cases} \tag{17}$$

On the condition of $v_k > v_{tra}-C$ and $\theta>0$, (16) is guaranteed to be positive. Furthermore, combining the condition of $\theta>1$, (17) is ensured to be negative. Observed from (15), since x_k is guaranteed to be smaller than λ_k by (9), U_{mk} is positive on the premise of $v_k<v_{tra}+g$. According to the above analysis, provided that $v_{tra}-C<v_k<v_{tra}+g$ and $\theta>1$, the partial derivative of the utility of CP is negative definite matrix. Then the optimal caching incentive price v_k^* that maximizes the utility of CP is obtained by setting the first order derivative of W_s as 0 as shown in (18). We denote the optimal incentive caching price vector as $V^*=\{v_1^*,v_2^*,...,v_K^*\}$.

$$\frac{\partial W_s}{\partial v_k^*} = 0 \tag{18}$$

Therefore, we can conclude that given the condition of $v_{tra}-C<v_k<v_{tra}+g$ and $\theta>1$, the Stackelberg equilibrium of the game between ISP and CP exists, denoted as (X^*,V^*). At Stackelberg equilibrium, CP and ISP can obtain their maximum utility simultaneously [15].

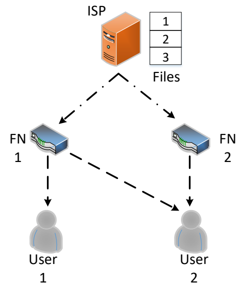
B. Interaction between users and FNs

With the global caching amount decided, reducing the average download delay is the only way to optimize the utility of users as we demonstrated in problem formulation. Therefore, after the global caching amount of each class of content is solved in the interaction between ISP and CP as we described above, we allocate data to be cached in various FNs to minimize the average download delay of users. The problem we expect to solve in this section is which FN to cache which data. The optimization function is shown as (10), which is in the form of the joint probability distribution of variables h_{lk}. The problem of determining the caching strategy $H=\{h_{lk}\}$ is to solve the marginal probability distribution problem under the minimum average download delay condition, which is NP-hard [10]. Belief propagation is one way to solve marginal probability distribution problem in probability graph and can get a suboptimal solution in a distributed way [16]. As FN only has local information such as local users requests distribution, belief propagation is suitable for solving this cache allocation problem. We demonstrate the design details of our belief propagation based cache placement algorithm by first present the factor graph model and then clarify the message passing procedure.

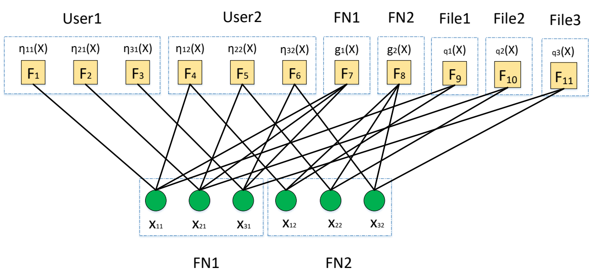
1) Factor Graph Model: In order to make use of belief propagation algorithm, we first convert the optimization function from the form of polynomials sum to the form of polynomials product and present the mapping rule of constructing factor graph. An explanatory example of transforming physical network connection to factor graph model with 2 FNs, 2 users and 3 classes of data is given in Fig.2.

Let $\eta_{gk}(H)=exp(-p_{gk}\overline{D}_{gk}(H))$ represents the download delay of transmitting a chunk of data f_k to user u_g, binary variable $g_l(H)$ denotes the cache capacity constraint of FN n_l and binary variable $q_k(H)$ denotes the global caching amount decision derived from Stackelberg game which stand for the first and second constraint in (11) respectively. $g_l(H)=1$ if $\sum_{k=1}^{K}h_{lk}\leqslant A_l$, otherwise $g_l(H)=0$. $q_k(H)=1$ if $\sum_{l=1}^{L}h_{lk}\leqslant x_k$, otherwise $q_k(H)=0$. Then the optimization function (10) can be converted from the form of polynomials sum to the form of polynomials product and from a minimization problem to a maximization problem as shown below:

$$\widehat{H} = arg \max_{h_{lk}\in\{0,1\}} \prod_{g\in[1,G],k\in[1,K]} \eta_{gk}(H)\prod_{l=1}^{L}g_l(H)\prod_{k=1}^{K}q_k(H) \tag{19}$$



(a) physical connection



(b) factor graph

Fig. 2. An explanatory example of transforming physical network connection to factor graph model

Each element of H corresponds to a variable node in factor graph denoted as s_i and each function $\eta_{gk}(H)$, $g_l(H)$ or $q_k(H)$ corresponds to a function node in factor graph denoted as F_j. The mapping rules are presented in equation (20) and (21). The index i of variable h_{lk} in all variable nodes is determined by its position in the matrix $H = \{h_{lk}\}$ following the row order. The index j of function $\eta_{gk}(H)$ in all function nodes is constructed in the same way as index i. The rest of the function nodes is constructed by placing functions $g_l(H)$ after $q_k(H)$ sequentially.

$$s_i = h_{lk}, \quad i = (l-1)K + k \tag{20}$$

$$F_j = \begin{cases} \eta_{gk}(H), & j = (g-1)K + k \\ g_l(H), & j = GK + l \\ q_k(H), & j = GK + L + k \end{cases} \tag{21}$$

The connection between variable nodes and function nodes in factor graph implicates both the physical connection between users and FNs and the requests connection between users and various classes of data. For each index k, variable nodes $s_i = h_{lk}$ connect to function nodes $F_j = \eta_{gk}(H)$ on the condition that user u_g is in the coverage range of FN n_l. Variable node $s_i = h_{lk}$ is in connection with function nodes $F_j = g_l(H)$ or $F_j = q_k(H)$ that has the same index l or k with h_{lk}. Let Γ_i^s denotes the neighboring function nodes of variable node s_i and Γ_j^F denotes the neighboring variable nodes of function node F_j.

2) Message Passing Procedure: In each iteration, message is passed between adjacent variable node and function node to exchange belief for the variable s_i. The belief for variable s_i is a value that indicates whether variable $s_i = h_{lk}$ should be 0 or 1. Variable $s_i = 1$ if the belief for s_i is greater than 0. Let $\alpha_{i\to j}^t$ and $\beta_{j\to i}^t$ denotes the message passing from variable node s_i to function node F_j and message passing from function node F_j to variable node s_i in the tth iteration respectively. The update rules of $\alpha_{i\to j}^t$ and $\beta_{j\to i}^t$ are presented as follows:

$$\alpha_{i\to j}^t = \sum_{l\in\Gamma_i^s\setminus\{j\}} \beta_{l\to i}^t \tag{22}$$

$$\beta_{j\to i}^t = \begin{cases} p_{gk}(\overline{D}_{gk}(H_{i,0}^t) - \overline{D}_{gk}(H_{i,1}^t)), & F_j = \eta_{gk}(H) \\ min\{0, -\alpha_{e\to j}^{(A_l)}(t)\}, & F_j = g_l(H), \quad e \in \Gamma_j^F \setminus \{i\} \\ min\{0, -\alpha_{e\to j}^{(x_k)}(t)\}, & F_j = q_k(H), \quad e \in \Gamma_j^F \setminus \{i\} \end{cases} \tag{23}$$

where the elements of $H_{i,0}^t$ and $H_{i,1}^t$ are binary variables and set as $h_{lk} = s_q = 1(q \in E_i^t = \{i_1 \in \Gamma_j^F \setminus \{i\} | \alpha_{i_1\to j}^t > 0\})$ and $h_{lk} = s_q = 1(q \in E_i^t \cup \{i\})$ respectively. By setting s_i in $H_{i,0}^t$ and $H_{i,1}^t$ as 0 and 1 respectively while keeping the value of other variable nodes in $H_{i,0}^t$ and $H_{i,1}^t$ the same, the updating rule of message $\beta_{j\to i}^t$ for function node $F_j = \eta_{gk}(H)$ indicates the delay gap in each case. $\alpha_{e\to j}^{(A_l)}(t)$ and $\alpha_{e\to j}^{(x_k)}(t)$ represents the A_lth and x_kth message among the messages $\{\alpha_{e\to j}^t\}(e \in \Gamma_j^F \setminus \{i\})$ arranged in the descending order respectively. Noted that the message from function node $F_j = g_l(H)$ and $F_j = q_k(H)$ to variable nodes can not be greater than zero. Take the case of $\beta_{j\to i^*}^t$ from $F_j = g_l(H)$ to variable node s_{i^*} for example, if $\alpha_{e\to j}^{(A_l)}(t) > 0(e \in \Gamma_j^F \setminus \{i^*\})$ which means that except variable node s_i^*, at least A_l neighboring variable nodes of function node $F_j = g_l(H)$ decide to take its value as 1. Then variable node s_i^* should not take its value as 1 in case of violating the FN cache capacity constraint. Therefore, $\beta_{j\to i^*}^t$ for function node $F_j = g_l(H)$ should not be greater than zero, because $\alpha_{i^*\to j}^t$ is updated as the sum of the messages from its neighboring function nodes as shown in (22). The design of the updating rule of $\beta_{j\to i}^t$ for function node $F_j = q_k(H)$ reflects similar design philosophy as the updating rule of $\beta_{j\to i}^t$ for function node $F_j = g_l(H)$.

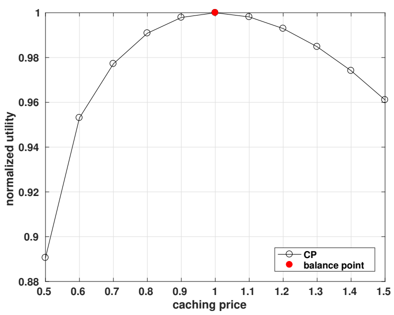
The belief for each variable $s_i = h_{lk}$ is updated in each iteration and is obtained as:

$$b_i^t = \sum_{j\in\Gamma_i^s} \beta_{j\to i}^t \tag{24}$$

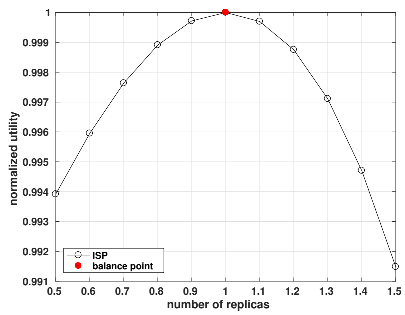
According to the belief b_i^t, we can estimate s_i as 1 if the corresponding belief b_i^t is greater than 0, otherwise estimate s_i as 0. Convergence of the estimated s_i leads to the termination of the iteration procedure.

VI. Simulation Results and Analysis

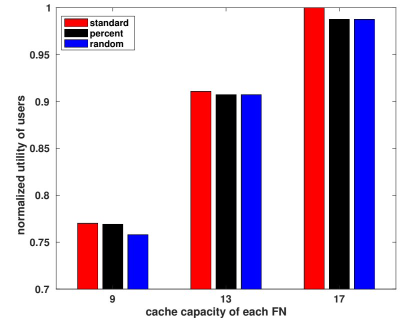
In this section, we present the simulation results and mainly evaluate the following: 1) the joint optimization of utilities of CP, ISP and users at Stackelberg equilibrium (X^*, V^*); 2) the effectiveness of the incentive caching scheme on shaping the pricing and caching behavior of CP and ISP respectively;



(a) utility of CP

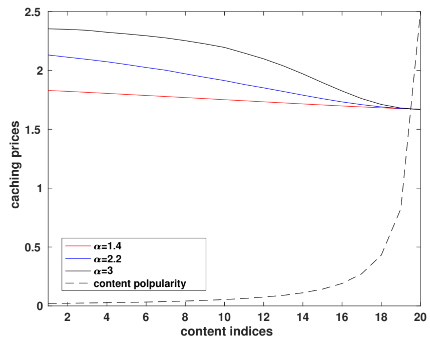


(b) utility of ISP

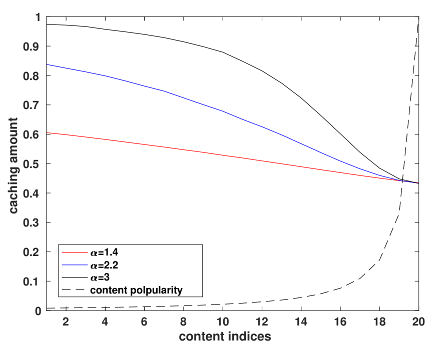


(c) utility of users

Fig. 3. Stackelberg equilibrium

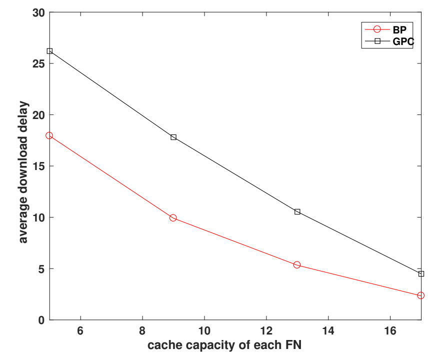


(a) pricing behavior of CP

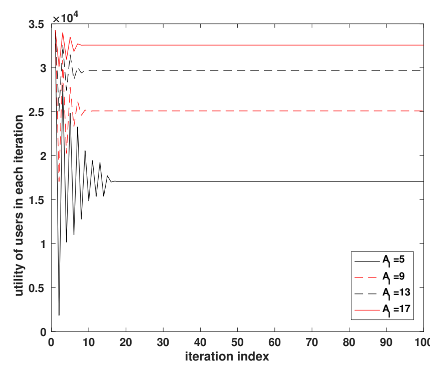


(b) caching behavior of ISP

Fig. 4. the impact of incentive caching scheme on shaping the pricing and caching behavior of CP and ISP



(a) download delay vs. cache capacity



(b) iteration procedure

Fig. 5. Performance of belief propagation based cache placement scheme

3) the performance of belief propagation based cache placement strategy on convergence speed and reducing the average download delay of users.

We consider an fog computing network with one ISP, one CP, 7 FNs and 50 users. The distance between two FNs is 200m and FNs provide caching service for users that locate within the radius of 150m. Users are randomly distributed and make 2000 requests in total for 20 classes of content following Zipf distribution with the parameter α. Larger α means more requests are distributed on smaller portion of popular content. The unit cost of ISP is set as $C_0 = 0.3$ and $C = 0.5$. The unit transmission fee v_{tra} and content price v_{con} is set as 2 and 3 respectively. We also set $\zeta_{gk} = 40$, $\theta = 2$ and $g = 8$. We reference [17] for the calculation method of $\overline{D}_{gk}(H)$.

In Fig.3, we demonstrate the joint optimization of utilities of CP, ISP and users at Stackelberg equilibrium. Setting α as 1.6, we calculate the Stackelberg equilibrium (X^*, V^*) denoted as filled red dot in Fig.3(a), Fig.3(b) and red bar in Fig.3(c) according to the method elaborated in Section IV. Deviation from V^* leads to decline in the utility of CP as shown in Fig.3(a). It is observed from Fig.3(b) that ISP fails to obtain the maximum utility if its global caching amount decision violates equation (13). Furthermore, in contrast to set global caching amount $X = \{x_1, x_2, ..., x_K\}$ according to requests distribution and as random permutation of $X^* = \{x_1^*, x_2^*, ..., x_K^*\}$ respectively which is denoted as black and blue bars in Fig.3(c), X^* optimizes the utility of users. Therefore, we can draw a conclusion that the utilities of CP, ISP and users reach a joint optimization state at Stackelberg equilibrium (X^*, V^*).

As shown in Fig.4, the effectiveness of the incentive caching scheme on shaping the pricing and caching behavior of CP and ISP is evaluated. We plot in Fig.4(a) the incentive caching prices offered by CP for content with different degrees of popularity. From the perspective of ISP, it prefers to cache popular content for the purpose of achieving higher cache hit rate. It is observed that CP offers higher incentive caching price for those less popular content. With larger α, the pricing difference is more distinct. Normalized by the number of requests for corresponding classes of content, the global caching amount for different classes of content decided by ISP is plotted in Fig.4(b). Although ISP allocates larger amount of cache to prevalent content, larger portion of requests for less prevalent content is satisfied by the caching replicas. Similarly, the gap in caching amount normalized by corresponding requests amount among different classes of content widen as α increases.

The performance of our proposed belief propagation based cache placement strategy is demonstrated in Fig.5. Fig.5(a) plots user's average download delay in the case of different cache capacity of FNs when applying our proposed belief propagation based cache placement strategy. The contrast schemes are global popularity based caching strategy which caches the most popular contents based on the statistical preference of users in the whole network. Our proposed algorithm can lower users' average download delay by up to 33.7% compared with global popularity based one. Average download delay drops as cache capacity enhances. That's because with larger cache capacity, FNs can cache more content and more requests are satisfied by FNs. In Fig.5(b), we analyze the converge speed of our proposed belief propagation based cache placement strategy in the case of various caching capacities of FNs. It is shown that within tens of iterations, the utility of users converges to a constant and optimal value, which indicates the feasibility of our proposed caching scheme.

VII. Conclusion and Future Work

In this paper, we propose a joint utility optimization caching resource allocation strategy for fog computing scenario combining global and local decision making. Utility functions of CP, ISP and users in our considered fog computing scenario are defined separately and considered as our optimization objective function. Given the selfishness nature of CP and ISP, we argue the necessity of considering the benefits of all network participants simultaneously in the design of caching scheme. We consider the interaction between CP and ISP as a non-cooperative hierarchy Stackelberg game and mathematically prove the existence of Stackelberg equilibrium, where all network participants achieve their maximum utility. In the interaction between FNs and users, a belief propagation based cache placement algorithm is proposed to further optimize the utility of users. Simulation results prove the effectiveness of our proposed caching scheme in optimizing the utility of CP, ISP and users simultaneously.

Our future work focuses on extending the system model to accommodate multiple CPs, multiple ISPs and multiple users scenario, thus enhancing the applicability of our proposed caching scheme. We also plan to investigate the Stackelberg equilibrium between multiple CPs and multiple ISPs.

References

[1] S. Sarkar and S. Misra, "Theoretical modelling of fog computing: a green computing paradigm to support iot applications," *Iet Networks*, vol. 5, no. 2, pp. 23–29, 2016.

[2] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Edition of the Mcc Workshop on Mobile Cloud Computing*, 2012, pp. 13–16.

[3] R. Mahmud, R. Kotagiri, and R. Buyya, "Fog computing: A taxonomy, survey and future directions," in *Internet of everything*, 2018, pp. 103–130.

[4] H. Chen, Q. Chen, R. Chai, and D. Zhao, "Utility function optimization based joint user association and content placement in heterogeneous networks," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2017, pp. 1–6.

[5] S. Wang, X. Huang, Y. Liu, and R. Yu, "Cachinmobile: An energy-efficient users caching scheme for fog computing," in *Ieee/cic International Conference on Communications in China*, 2016, pp. 1–6.

[6] T. Liu, J. Li, B. Kim, C.-W. Lin, S. Shiraishi, J. Xie, and Z. Han, "Distributed file allocation using matching game in mobile fog-caching service network," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 499–504.

[7] X. Cao, J. Zhang, and H. V. Poor, "An optimal auction mechanism for mobile edge caching," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 388–399.

[8] A. Kiani and N. Ansari, "Toward hierarchical mobile edge computing: An auction-based profit maximization approach," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2082–2091, 2017.

[9] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2115–2129, 2016.

[10] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[11] C. Guo, G. Lu, H. J. Wang, S. Yang, C. Kong, P. Sun, W. Wu, and Y. Zhang, "Secondnet: A data center network virtualization architecture with bandwidth guarantees," in *Proceedings of the 6th International COnference*, ser. Co-NEXT '10. New York, NY, USA: ACM, 2010, pp. 15:1–15:12.

[12] J. Ferdous, M. P. Mollah, M. A. Razzaque, M. M. Hassan, A. Alamri, G. Fortino, and M. C. Zhou, "Optimal dynamic pricing for trading-off user utility and operator profit in smart grid," *IEEE Transactions on Systems Man and Cybernetics Systems*, vol. PP, no. 99, pp. 1–13, 2017.

[13] Y. Xu, Y. Li, C. Song, T. Lin, and F. Chen, "Distributed caching via rewarding: An incentive caching model for icn," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6.

[14] J. Stewart, "Multivariable calculus : concepts and contexts," 2010.

[15] E. Rasmusen, "Games and information : an introduction to game theory," *St.ewi.tudelft.nl*, vol. 9, no. 3, pp. 841–846, 1989.

[16] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2002.

[17] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *ICC 2016 - 2016 IEEE International Conference on Communications*, 2016, pp. 1–6.